



Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at <http://about.jstor.org/participate-jstor/individuals/early-journal-content>.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

ON THE THEORY OF CORRELATION WITH SPECIAL REFERENCE TO CERTAIN SIGNIFICANT LOCI ON THE PLANE OF DIS- TRIBUTION IN THE CASE OF NORMAL CORRELATION.

By H. L. RIETZ.

1. **Introduction.**—The notion of correlation is of such importance in science that it seems it should become almost as familiar to the scientist as the notions of a mathematical function and of independence in the probability sense. The purposes of the present paper are (1) to present the elements of a theory of correlation from assumptions that are suggested by applications and that seem to appeal to the mathematical student beginning the study of statistics, (2) to give a few properties of normally correlated statistical data by means of curves or contour lines on what I call the plane of distribution.

Take X and Y to represent associated classes of individuals that have definite values or with attributes that have definite values. These classes may represent any one of a great variety of concrete situations. To illustrate, X may represent rainfalls at a given place in months of April and Y those in months of June; X and Y may refer to fathers and sons, when the inheritance of some character is in question; X and Y may represent statures of husbands and of their wives; X may represent hours per day worked by laborer, and Y the corresponding wages paid. These illustrations very naturally suggest the following questions that show the purpose of a theory of correlation: Is there a measurable tendency for wet Aprils to be followed by dry Junes? To what extent do a class of men, in general, resemble their father with respect to some character, say stature? Do tall men, in general, marry tall wives? Do high wages go with short hours of labor?

The problem that we set is to describe, by some summary method, the tendency of corresponding individuals of X and Y to vary in the same or in opposite directions, when the variations are to be attributed to an indefinite number of unassignable causes. It is well known that the relationship of corresponding individuals of X and Y , in the illustrations cited above, is not such a perfect dependence as is given by a mathematical function. When a value is assigned to X , the corresponding values of Y have a certain amount of freedom. They may, however, be far from free in the probability sense of freedom or independence.

A relation expressed by a mathematical function and independence in the probability sense may be regarded as two extremes between which there exists a large region for a theory of correlation.

2. Table of Double Classification.—To be precise, let

$$x_1, x_2, x_3, \dots, x_n$$

be values of the individuals of a random sample of n drawn from the class X . In the language of statistics, we call these x 's variates. Let y 's represent the corresponding variates of the class Y . That is to say, X and Y are associated so that

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

are pairs of variates in correspondence.

The first step in the description of the dependence of the two attributes is the construction of a table* of double classification of the following form:

	X_1	X_2	X_3				X_s					X_i	
Y_m	N_{1m}	N_{2m}	N_{3m}				N_{sm}					N_{im}	N_{ym}
Y_t	N_{1t}	N_{2t}	N_{3t}				N_{st}					N_{it}	N_{yt}
$\vdots$													
Y_3	N_{13}	N_{23}	N_{33}				N_{s3}					N_{i3}	N_{y3}
Y_2	N_{12}	N_{22}	N_{32}				N_{s2}					N_{i2}	N_{y2}
Y_1	N_{11}	N_{21}	N_{31}				N_{s1}					N_{i1}	N_{y1}
	N_{z1}	N_{z2}	N_{z3}				N_{zs}					N_{zi}	n

FIG. 1.

The symbols $X_1, X_2, \dots$ in the top row of the table mark subclasses covering, in general, equal intervals on the range that includes the entire class X . Similarly, $Y_1, Y_2, \dots$ in the column on the left mark subclasses that are taken to include, in general, equal intervals of the range that includes the entire class Y . The number of such subclasses may be two or more.

In this table (Fig. 1), any number, say N_{st} ($s \neq x, s \neq y$), indicates the number that belongs to both subclasses X_s and Y_t . The vertical column of frequencies $N_{s1}, N_{s2}, N_{s3}, \dots, N_{st}, \dots$ corresponding to any mark X_s is called an X -array of y 's. Similarly, any row of frequencies $N_{1t}, N_{2t}, N_{3t}, \dots, N_{it}, \dots$ corresponding to any mark Y_t is a y -array of x 's.

The numbers in the lower row give the sums of numbers N in columns, and show the frequency distribution of the total sample of n into subclasses $X_1, X_2, \dots$. Likewise, the column of totals on the right is the frequency distribution into subclasses $Y_1, Y_2, \dots$.

* Such a table used to study the correlation of statistical data seems to have been employed first by Francis Galton, Proc. Royal Society, XL, p. 68.

3. Test of Independence.—If the two classes X and Y are independent in the probability sense, the best estimate (from our sample of n) that can be given, from the separate classes, of the probability that a pair (x_p, y_p) taken at random is such that x_p belongs to the subclass X_s and y_p to Y_t , is the product

$$\frac{N_{xs} \cdot N_{yt}}{n^2}. \quad (1)$$

But, from the table (Fig. 1), the best estimate that a pair belongs to both these classes is N_{st}/n . If the deviations

$$\delta = \frac{N_{st}}{n} - \frac{N_{xs} \cdot N_{yt}}{n^2}, \quad (2)$$

when extended to all compartments of the table, are greater than may be attributed to fluctuations in drawing random samples, there is lack of independence. If the differences (2) are to be attributed to smallness of the sample, we say the difference is insignificant. A test as to whether the deviations δ are to be regarded as significant has been given by Sheppard.*

If the classes X and Y are not independent, and are not absolutely dependent in the sense that a mathematical function determines the one when the other is given; then, we may say, in a general way, that a theory of correlation is required to describe the association of the two classes X and Y .

4. General Description.—In many applications, where the correlation is considerable, a certain amount of progress in treating the association of X and Y can be made by a consideration of the arrays of the double entry table of values without the use of special mathematical methods. That is to say, we may treat each array as a frequency distribution, and find some kind of average values for the arrays and for the variability from this average. Such a treatment is, however, obviously inadequate for many purposes.

Suppose we erect at the center of each rectangle (Fig. 1) of the table a perpendicular to the plane of the rectangle proportional in length to the number in that rectangle. The ends of these perpendiculars suggest a surface to describe the distribution.

5. Geometrical Description.—To follow the suggestion just mentioned, it is convenient to use $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ as deviations from the means of variates of the respective classes X and Y rather than for the actual values of the variates themselves. They will be used in this sense throughout the remainder of the paper.

* Philosophical Transactions of the Royal Society, vol. 192A, 1899, p. 128.

Represent these deviations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from the mean values as coördinates of points in a plane (Fig. 2*). I call this plane the *plane of distribution*. It is the purpose of a theory of correlation to characterize the arrangement of such points without special reference to individual points. Divide the range along the x -axis that would include all points of the class into equal intervals Δx . Likewise, divide the range along the Y -axis into equal intervals Δy . The points included by two parallels to the Y -axis (AB) may be said to constitute an x -array of Y 's. Similarly, the points included by parallels ($A'B'$) are said to be a Y -array of x 's.

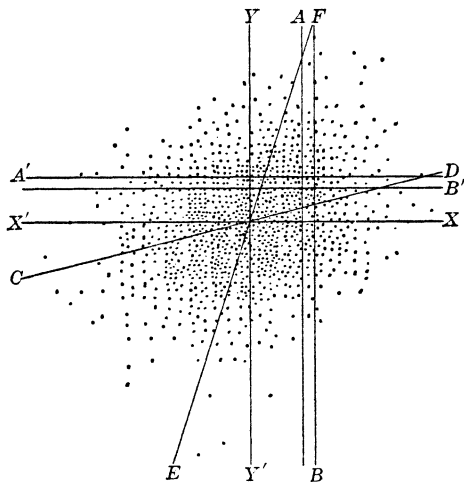


FIG. 2.

It is reasonable to assume that some function $f(x)$ exists such that

$$f(x)\Delta x \quad (3)$$

is the probability that a variate taken at random from class X gives a point in an array marked by a Δx , and that some function $\varphi(y)$ is such that

$$\varphi(y)\Delta y \quad (4)$$

is the probability that a variate taken at random from Y gives a point in an array marked Δy . The table of double classification (Fig. 1) with perpendiculars to the plane at the centers of rectangles suggested the idea of a surface. To follow this suggestion, we consider an area $\Delta x \Delta y$ at the intersection of any two arrays marked Δx and Δy . It may be assumed that

* This distribution of points is not made up in an entirely artificial manner, but represents approximately the distribution of a class of husbands and wives with respect to stature. See Pearson and Lee, *Biometrika*, vol. 2, p. 408.

a function $z = \psi(x, y)$ exists such that

$$\psi(x, y)\Delta x\Delta y \quad (5)$$

gives the probability that a pair (x_p, y_p) taken at random gives a point in $\Delta x\Delta y$.

We may obviously add the conditions

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \quad \int_{-\infty}^{+\infty} \varphi(y)dy = 1, \quad \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \psi(x, y)dxdy = 1, \quad (6)$$

if the functions in (3), (4), and (5) are integrable from $-\infty$ to $+\infty$.

To render more precise and useful the idea of independence discussed in §3, we may now use the notation of §5, and define that the two classes X and Y are independent if

$$\psi(x, y) = f(x)\varphi(y) \quad (7)$$

is an identity. If this equality is not an identity, the two classes are correlated. From this general negative definition of correlation, it seems difficult to determine the character of $\psi(x, y)$ so as to give it a practical value. We seek therefore an affirmative definition at some loss of generality. To obtain a valuable affirmative definition, we fix our attention on the mean values of those variates of one class, say of Y , that are in an array marked Δx . Or, we may regard the points (Fig. 2) as particles of equal mass and fix our attention on the centroid of these particles. With the limiting case of an indefinitely large number of variates and small values of Δx , it seems reasonable, since mean values have variations of a higher order of smallness than individuals, that these centroids, in general, arrange themselves along a smooth curve. That is to say, there may be a correspondence, given by a mathematical function, between any assigned X and the centroid of corresponding points of Y .

With reference to the surface

$$z = \psi(x, y),$$

this means that the y coördinate, $\bar{y}$, of the centroid of any section by assigning x is a function of x . Suppose that

$$\bar{y} = \frac{\int_{-\infty}^{+\infty} y\psi(x, y)dy}{\int_{-\infty}^{+\infty} \psi(x, y)dy} = \theta(x) \quad (8)$$

for any assigned x .

As an affirmative, though special definition of correlation, we define that the class Y is correlated with X if $\theta(x)$ is different from zero. As we have

selected the origin so that $\theta(x)$ cannot be a constant different from zero, the simplest case of correlation we have to consider is that for which

$$\theta(x) = mx + b, \quad (9)$$

a linear function of x .

It affords a vivid description to interpret $\theta(x)$ on the plane of distribution (Fig. 2). It is sometimes called the curve of regression of y 's on x 's. Moreover, when $\theta(x)$ is linear, it is said that there is linear regression. Obviously, another curve of regression of x 's on y 's exists.

To obtain the m and b to apply to any numerical case, we make use of our sample of n . If the line (9) is subjected to the least squares condition that m and b are to be determined so that the sum of the squares of its deviations (measured parallel to the y -axis) from the means of arrays (weighted with the number in the arrays) is to be a minimum, we obtain

$$y = r \frac{\sigma_y}{\sigma_x} x,^* \quad (10)$$

where σ_x^2 is the mean square of $x_1, x_2, \dots, x_n$,

σ_y^2 is the mean square of $y_1, y_2, \dots, y_n$,

and r is defined by the formula

$$r = \frac{\sum_{q=1}^{q=n} x_q y_q}{n \sigma_x \sigma_y}.$$

The σ 's are called the "standard deviations" of the systems of variates and r is called the "correlation coefficient."

The line (10) is the line of distribution of the means of y 's that correspond to assigned x 's. By analogy,

$$x = r \frac{\sigma_x}{\sigma_y} y \quad (11)$$

is the line of regression of x 's on y 's. From our special definition, there is no correlation if $r = 0$.

7. Standard Deviation of Arrays.—The mean square of the deviations of all points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, [Fig. 2], from corresponding points

$$\left(x_1, r \frac{\sigma_y}{\sigma_x} x_1\right), \left(x_2, r \frac{\sigma_y}{\sigma_x} x_2\right), \dots, \left(x_n, r \frac{\sigma_y}{\sigma_x} x_n\right)$$

on the line

$$y = r \frac{\sigma_y}{\sigma_x} x$$

is given by

* See Fig. 2. Line CD is the line of regression of wives with respect to husbands, and EF is the line of regression of husbands with respect to wives.

$$\begin{aligned}
 \frac{\sum_{q=1}^{q=n} \left(y_q - r \frac{\sigma_y}{\sigma_x} x_q \right)^2}{n} &= \frac{\sum_{q=1}^{q=n} y_q^2}{n} - \frac{2\sigma_y r}{\sigma_x} \frac{\sum_{q=1}^{q=n} x_q y_q}{n} + \frac{r^2 \sigma_y^2}{\sigma_x^2} \frac{\sum_{q=1}^{q=n} x_q^2}{n} \\
 &= \sigma_y^2 - 2r^2 \sigma_y^2 + r^2 \sigma_y^2 \\
 &= \sigma_y^2 (1 - r^2).
 \end{aligned} \tag{12}$$

Suppose that the regression is linear, so that the centroid of the x -arrays of y 's may be taken on the line

$$y = r \frac{\sigma_y}{\sigma_x} x,$$

and that the standard deviations of x -arrays of y 's are equal. Then $\sigma_y^2(1 - r^2)$ becomes the square of the standard deviation of each x -array of y 's. When the standard deviations of these arrays are unequal, $\sigma_y^2(1 - r^2)$ is merely a sort of an average value of the square of standard deviations of arrays.

Similarly, $\sigma_x^2(1 - r^2)$ is the square of the standard deviation of a y -array of x 's.

8. The Normal Correlation Surface.—The normal correlation surface dates back to a memoir by Bravais* in 1846. The importance of this surface in the mathematics of statistics was first recognized by Galton† and has been fully demonstrated by the work of Pearson,‡ Edgeworth, and Yule.

The equation of the normal surface may be written in the form

$$z = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y} \right)}. \tag{13}$$

For certain associated classes X and Y , this is the form of $\psi(x, y)$ [§ 5, (5)]. When $\psi(x, y)$ takes this form, the correlation is said to be normal.

It is easy to give many properties of this surface from certain sets of which equation (13) may be derived. One set of conditions that characterize the surface and that suggest themselves very naturally in attempting a description of certain double entry tables of classification may be stated as follows:

(a) The regression of at least one set of variates on the other is linear. Applied to the surface, this condition means that

$$\theta(x) = r \frac{\sigma_y}{\sigma_x} x. \tag{[§ 6, (8)]}$$

* Sur les Probabilités des Erreurs de Situation d'un Point. Memoires par divers Savants a l'Académie des Sciences de France, t. IX (1846), pp. 255-332.

† Galton, Proc. Roy. Soc., vol. XL, p. 42 (1886).

‡ Pearson, Phil. Trans. (A), vol. 187, p. 253 (1896); A, vol. 200, p. 1 (1903). Edgeworth, Phil. Mag., 1892, vol. 34, p. 190 (1903). Yule, Journal of the Royal Statistical Society, vol. 60, p. 812. Cf. Bertrand, Calcul de Probabilités, Chap. IX, Czuber, Theorie der Beobachtungsfehler, Dritter Teil.

(b) The arrays of the correlation table are normal distributions; that is, any section of the surface $z = \psi(x, y)$ by assigning x or y is a normal* (Gaussian) curve.

These conditions, if we make an additional assumption of a purely analytic nature, lead so readily to (13) that we shall not give the details of the argument.

Condition (b) implies that $\psi(x, y)$ is of the form $e^{X(x, y)}$, and the additional assumption is that $X(x, y)$ can be expanded in a convergent power series in x and y .

Certain Loci on the Plane of Distribution in Normal Correlation.

9. **Ellipses of Equal Probability.**—In what follows, it seems to me to add somewhat to the description to regard points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (Figs. 2 and 3) as particles of equal mass on the plane of distribution. The average density of distribution of particles on an area may be defined as the relative frequency† of particles on that area divided by the area. From this definition, we may pass to the limiting case and say that z in the surface

$$z = \psi(x, y)$$

gives the limiting value of the density at any point on the plane of distribution. The curve along which the density of distribution is constant is the ellipse obtained by assigning a given value to z in equation (13), and interpreting the result on the plane of distribution. The infinite system of homothetic ellipses obtained by assigning different values to z plays an important role in Bravais's fundamental memoir.‡ Such an ellipse is sometimes called an ellipse of equal probability.§ We shall deal in this paper (§ 10) with one of these ellipses of special interest. For this purpose, the equation of any ellipse of the system may be written in the form

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y} = \lambda^2. \quad (18)$$

The area of this ellipse is

$$\frac{(\pi\lambda^2\sigma_x\sigma_y)}{\sqrt{1-r^2}}$$

and the semiaxes are given by $a = k\lambda$ (19) and $b = k'\lambda$ (20), where k and

* The normal curve in rectangular coordinates is defined by the equation

$$y = e^{ax^2+bx+c},$$

where a is negative. It is easily shown that $a = -1/2\sigma^2$, where σ^2 is defined as the second moment of the area under the curve, about the line $x = -b/a$, divided by the area.

† The "relative frequency" of events and the probability of an event are used interchangeably in this paper.

‡ Loc. cit.

§ Cf. Bertrand, loc. cit.

k' are functions of σ_x , σ_y , and r . The probability that a particle will fall within any ellipse obtained by assigning λ , is given by

$$\frac{2\pi\sigma_x\sigma_y}{2\pi\sigma_x\sigma_y(1-r^2)} \int_0^\lambda e^{-\frac{1}{2(1-r^2)}\lambda^2} \lambda d\lambda = 1 - e^{-\frac{\lambda^2}{2(1-r^2)}}. \quad (21)$$

10. The Ellipse of Maximum Probability.—We shall now determine the ellipse along which, for a given small ring $\Delta\lambda$, we should expect more particles than along any other ellipse of the system.

The perimeter of the ellipse of semiaxes $k\lambda$ and $k'\lambda$ (§ 9) is given by

$$4k\lambda \int_0^{\frac{\pi}{2}} \sqrt{1 - e^2 \sin^2 \varphi} d\varphi,$$

where e^2 is independent of λ . Since the integral is independent of λ , we may write the perimeter of the ellipse in the form $k''\lambda$, where k'' is independent of λ . Hence, the total probability that a particle falls in a small ring between ellipses λ and $\lambda + \Delta\lambda$ is given by

$$k''\lambda e^{-\frac{\lambda^2}{2(1-r^2)}} \Delta\lambda. \quad (22)$$

This expression is a maximum when $\lambda^2 = 1 - r^2$. Hence, what may well be called the ellipse of maximum probability is

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y} = 1 - r^2. \quad (23)$$

To illustrate the meaning of this ellipse, in Bertrand's illustration of shooting a thousand shots at a target, the probability is greater that a shot will fall along this ellipse than along any other ellipse of the infinite system.

We may further easily prove the following theorem: *The ellipse of maximum probability is identical to the orthogonal projection of parabolic points of the correlation surface on the plane of distribution.*

To prove this theorem, we simply find the locus of parabolic points on the surface (13) by means of the well-known condition

$$\frac{\partial^2 z}{\partial x^2} \cdot \frac{\partial^2 z}{\partial y^2} = \left(\frac{\partial^2 z}{\partial x \partial y} \right)^2.$$

This gives

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y} = 1 - r^2,$$

which establishes the theorem.

Attention has often been called to another ellipse known as the "probable" ellipse. The probable ellipse is defined as that ellipse of the system

such that the probability is $\frac{1}{2}$ that a particle falls within it. This means, by (21), that λ is such that

$$e^{-\frac{\lambda^2}{2(1-r^2)}} = \frac{1}{2},$$

or

$$\lambda^2 = 1.3863(1 - r^2). \quad (24)$$

Hence, the probable ellipse is larger than the ellipse of maximum probability. In fact, the probability that a particle falls within the ellipse of maximum probability is $1 - e^{-1} = 0.3935$, while that of falling within the probable ellipse is, by definition, $\frac{1}{2}$.

For the illustration of statures of husbands and wives, these two ellipses are shown in Fig. 3. By actual count from the drawing (Fig. 3), it appears

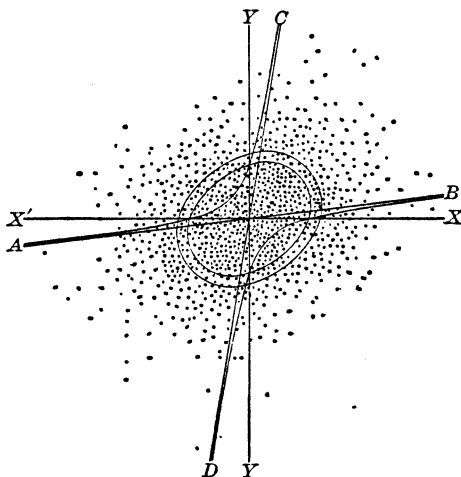


FIG. 3.

that 536 of the 1,078 points are within* the probable ellipse and 412 are within the ellipse of maximum probability. These numbers differ from the theoretical values by amounts well within what should be expected with 1,078 points in all.

11. Separation of the Plane of Distribution by Lines of Regression (Fig. 2).—The lines of regression and some other lines to be defined presently are of such importance in describing the distribution of particles on our plane of distribution that we shall consider the probability that a particle falls into a given compartment of the plane separated from the rest of the plane by these lines.

Let us take the lines

$$y = l_1x, \quad (25)$$

$$y = l_2x, \quad (26)$$

* One half the points on the ellipse are counted within it in making this count.

and determine the probability that a particle falls into a compartment made by these lines. The probability is given by the volume under the correlation surface bounded by (25) and (26), interpreted as planes. The probability is, if we make

$$x = \rho \cos \theta,$$

$$y = \rho \sin \theta,$$

$$P = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \int_{\arctan l_1}^{\arctan l_2} \int_0^\infty e^{-\frac{\rho^2}{2(1-r^2)}\left(\frac{\cos^2\theta}{\sigma_x^2} + \frac{\sin^2\theta}{\sigma_y^2} - \frac{2r\sin\theta\cos\theta}{\sigma_x\sigma_y}\right)} \rho d\rho d\theta$$

$$= \frac{1}{2\pi} \left(\arctan \frac{\sigma_x l_2 - r\sigma_y}{\sigma_y\sqrt{1-r^2}} - \arctan \frac{\sigma_x l_1 - r\sigma_y}{\sigma_y\sqrt{1-r^2}} \right). \quad (27)$$

If (25) and (26) are lines of regression so that

$$l_1 = r \frac{\sigma_y}{\sigma_x}, \quad l_2 = \frac{1}{r} \frac{\sigma_y}{\sigma_x},$$

then

$$P = \frac{1}{2\pi} \arccos r \quad (0 \leq \arccos r \leq \pi). \quad (28)$$

When r is positive, this is the probability that a point will fall into one of the compartments of the smaller angles between the lines of regression.

To find the probability that a particle falls into this same region under independence of X and Y , we make $r = 0$ in (27) before making the substitutions

$$l_1 = r \frac{\sigma_y}{\sigma_x}, \quad l_2 = \frac{1}{r} \frac{\sigma_y}{\sigma_x}.$$

This gives

$$P' = \frac{1}{2\pi} \arccos \frac{2r}{1+r^2}.$$

Hence, the excess relative frequency in this compartment of the table is

$$\frac{1}{2\pi} \left(\arccos r - \arccos \frac{2r}{1+r^2} \right).$$

12. Loci Along Which the Frequency of Particles Bears a Simple Relation to the Frequency under Independence.—If the equality [§ 6, (7)] $\psi(x, y) = f(x) \cdot \varphi(y)$ is not an identity, it may be interpreted as the curve in the plane of distribution along which particles are distributed with the same frequency as they would be under independence. This curve and its separation of the plane of distribution have been treated by Pearson* for the case of normal correlation.

* Drapers' Company Research Memoirs, Biometric Series, I, XIII.

That treatment is easily extended for normal correlation to find the curve along which points are k times as frequent as on the hypothesis of independence. I call k the *intensity of distribution* with respect to independence. In the general case, this curve has the equation

$$\psi(x, y) = kf(x) \varphi(y). \quad (30)$$

For normal correlation, (30) takes the form

$$\frac{1}{2\pi\sigma_x\sigma_y} \sqrt{1-r^2} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y}\right)} = \frac{k}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)},$$

which may be easily simplified to

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2xy}{r\sigma_x\sigma_y} = \frac{(1-r^2)}{-r^2} \log(k^2 - k^2r^2). \quad (31)$$

This hyperbola divides the plane of distribution into two regions in one of which the intensity of distribution is greater than k , while in the other it is less than k . By giving different values to k , we obtain an infinite system of hyperbolas. The case of

$$k^2 = \frac{1}{1-r^2}$$

is of special interest as it may be regarded as the intensity of distribution at the centroid of particles. Making

$$k = \frac{1}{\sqrt{1-r^2}}$$

in (31) causes the equation to degenerate into the two straight lines

$$y = \frac{1}{r} \frac{\sigma_y}{\sigma_x} x (1 \pm \sqrt{1-r^2}). \quad (32)$$

These lines are shown as lines AB and CD on Fig. 3. The hyperbola along which the frequency is the same as under independence is also shown on the same figure. The probability that a particle will fall into a specified one of the four compartments [made by lines (32)] of the plane of distribution is given by substitution for l_1 and l_2 in (27). This gives for the probability $\frac{1}{4}$. That is, the probability is just the same that a particle belongs to one of the four compartments as to any other.

The probability that a point belongs to the region of one of these compartments, in case of independence, is

$$\frac{1}{2\pi} \arccos r.$$

13. Separation of Plane by What Would Be Lines of Regression Under Independence.—These lines are simply our coördinate axes. The probability that a particle falls into a specified quadrant under independence is, of course, $\frac{1}{4}$. The probability that a particle will fall into the first quadrant, in the case of normal correlation, is given by (27) as

$$\frac{1}{4} + \frac{1}{2\pi} \arcsin r \quad \left(-\frac{\pi}{2} < \arcsin r < \frac{\pi}{2} \right).$$

And the probability of falling into the fourth quadrant is

$$\frac{1}{4} - \frac{1}{2\pi} \arcsin r \quad \left(-\frac{\pi}{2} < \arcsin r < \frac{\pi}{2} \right).$$

Hence, we may say, in the case of normal correlation, if we know only in regard to a variate of class X that it is above the mean, that the odds are $\frac{1}{4} + 1/2\pi \arcsin r$ to $\frac{1}{4} - 1/2\pi \arcsin r$ that the corresponding variate of class Y is above the mean.

In our example of the correlation of husbands and wives in stature, the numerical values of the odds are 0.2962 to 0.2038 or 3 to 2 approximately that the stature of the wife is above the mean if it is given that the husband is above the mean of husbands in stature.

The study of the separations of the plane of distribution by such lines as those here considered throws considerable light on the character of normally correlated statistical data.

UNIVERSITY OF ILLINOIS.